The Effects of Voice Pitch and Race on Perceived Leadership Ability and Threat

Keana Richards[1]

[1] University of Pennsylvania

Abstract

Negative stereotypes about Black people are pervasive in America. Specifically, there is a widespread stereotype that Black men are prone to crime and violence (Quillian & Pager, 2001). Recently, it has been argued that personal characteristics that indicate an individual is less threatening (i.e., disarming mechanisms) may reduce the salience of this stereotype and, in turn, reduce barriers to employment and economic advancement (Livingston & Pearce, 2009). In support of this argument, research has demonstrated that there are more baby-faced Black male CEOs than baby-faced White CEOs and that Black male CEOs with a baby-faced appearance are more successful than non-baby-faced Black CEOs (Livingston & Pearce, 2009). Voice pitch, which influences perceptions of threat potential and leadership ability (Klofstad, Anderson & Peters, 2012; Puts, Apicella, & Cardenas, 2012), is another personal characteristic that may serve as a disarming mechanism. We experimentally tested whether voice pitch differentially affects perceptions of leadership and threat in Black and White men using a within-subjects design involving over 500 participants recruited from an online market. As expected, men with lower-pitched voices were rated as more threatening and as better leaders, regardless of their race. Unexpectedly, we found a main effect of race on perceived leadership ability, where Black men were rated higher on leadership traits than White men. Overall, the findings do not suggest that Black men with lower-pitched voices were disadvantaged relative to their counterparts with higher-pitched voices. We discuss possible explanations for our findings, with implications for Black men in the workplace.

Keywords: voice pitch, race, leadership, threat, men, stereotypes

Word count: X

The Effects of Voice Pitch and Race on Perceived Leadership Ability and Threat

Introduction

American society is plagued by the stereotype that Black people engage in criminal activity and violence (Quillian & Pager, 2001; Welch, 2007), which heightens perceptions of threat from these individuals (Cottrell & Neuberg, 2005). For instance, Black men are perceived as substantially more threatening when they are taller than average compared to White men (Hester & Gray, 2018). Other research shows that Black men are perceived as physically larger (i.e., taller and heavier) and more capable of physical harm compared to White men (Wilson, Hugenberg, & Rule, 2017). Overall, this line of research suggests that Black men are more likely to be the targets of stereotypes about their capacity to threaten others' physical safety, even when they do not pose any threat.

These perceptions contribute to institutional racism, where these individuals face systematic barriers across several domains. For instance, Black people may face differential treatment within the criminal justice system, where they are more likely to be wrongly convicted and punished more harshly for crimes. In support of this argument, eyewitnesses are more likely to select a Black man with prototypically Black facial features (e.g., wide nose, thick lips) as the offender when presented in a line-up of suspects, even when the individual did not commit the crime (Knuycky, Kleider, & Cavrak, 2014). Also, men that look prototypically Black are more likely to be sentenced to death when they have been convicted of murdering a White victim compared to a Black victim (Eberhardt, Davies, Purdie-Vaughns, & Johnson, 2006). Finally, individuals are more likely to shoot unarmed Black men in first-person shooter tasks (Correll, Park, Judd, & Wittenbrink, 2007). Another domain that may be affected by these stereotypes is the marketplace, where it has been demonstrated across numerous studies that there is severe employment discrimination against non-Whites and that Black individuals have poorer market outcomes relative to their White counterparts (Ayres, Banaji, & Jolls, 2015; Doleac & Stein, 2013; Riach &

Rich, 2002), which may contribute to the enduring disparity in socioeconomic status (SES) between Black and White individuals (Hayward, Miles, Crimmins, & Yang, 2000). For example, Ayres et al. (2015) manipulated the skin color of the hands holding cards during a baseball card auction on eBay and found that Black sellers received fewer offers than Whites. On the occasions that Black sellers received offers, they were substantially lower than offers to White sellers. Researchers suggest that a lack of trust towards Black sellers contributes to these pattern of results (Doleac & Stein, 2013), which may be due to the prominent stereotype about their threateningness.

The persistence and pervasiveness of these targeted stereotypes about threat based upon group membership may be explained by human evolutionary history. Specifically, lethal conflict plagued many inter-group encounters (Bowles, 2009; Neuberg & Cottrell, 2008), increasing the salience of group membership and strengthening the association between group membership and perceived threat. In the racially heterogeneous environment of America today, where cues that may indicate group membership are especially conspicuous, people assess threat potential from others using these superficial cues (e.g., skin color), even when they do not accurately reflect an individuals' threat potential (Neuberg & Schaller, 2016).

People may also attend to other characteristics outside of group membership, like facial and vocal characteristics, when assessing threat potential. For instance, research suggests that humans rapidly and automatically categorize faces along two dimensions: perceived valence and dominance (Todorov, Said, Engell, & Oosterhof, 2008). The valence dimension reflects ratings of trustworthiness, while the dominance dimension maps onto ratings of dominance. Feelings of threat may be magnified when an individuals' facial features have high ratings on the dominance dimension and low ratings on the valence dimension (Oosterhof & Todorov, 2008), largely because people perceive these individuals as willing to (as induced by the valence dimension) and capable of (as induced by the dominance dimension) threatening others. With regards to vocal characteristics, there is

extensive empirical evidence that lower-pitched voices are associated with greater threat potential, which is reflected by arm strength and testosterone levels (Hodges-Simeon, Gurven, Puts, & Gaulin, 2014; Puts, Apicella, & Cardenas, 2012). Humans have androgen receptors in their vocal folds (Voelter et al., 2008), which are sensitive to peripubertal exposure to testosterone. With higher levels of testosterone applied to the vocal fold receptors, the vocal chords will thicken and vibrate more slowly (Harries, Hawkins, Hacking, & Hughes, 1998), which in turn produces a lower pitch. In this way, voice pitch serves as an honest signal of threat potential, which makes people especially likely to use voice pitch as an indicator of threat potential (Hodges-Simeon, Gurven, & Gaulin, 2015; Hodges-Simeon et al., 2014). Along these lines, men who have a lower-pitched voice are more likely to be perceived as dominant and untrustworthy (O'Connor & Barclay, 2017; Puts, 2010; Puts, Gaulin, & Verdolini, 2006).

Since we are more likely to perceive certain out-group members as threatening based upon different stereotypes that permeate throughout our culture, individuals within these groups that have disarming mechanisms (i.e., personal characteristics that reduce perceptions of threat) are more likely to be successful. This has been supported by previous research, where psychologists have shown that there are more baby-faced Black male CEO's than baby-faced White CEO's and that Black male CEO's with a baby-faced appearance are more likely to be successful (Livingston & Pearce, 2009). The researchers suggest that these individuals were perceived as more trustworthy in a social context dominated by out-group members (e.g., corporate America) and may be more successful than other individuals within their racial group because their facial features serve as a cue to their low threat potential, which improves their interpersonal interactions in that specific social context. In support of this argument, researchers show that other personal characteristics, like sexual orientation, can serve as a disarming mechanism for Black men in leadership positions, where gay Black men are rated as better leaders compared to single-minority men (i.e., gay men or Black men) (Wilson et al., 2017). The researchers

suggest that these effects can be explained by a stereotype that gay men are less masculine, which reduces perceived threat.

Perceptions of leadership ability are likely to be affected by stereotypes and personal characteristics because people do not want leaders that they perceive as a threat to their group, as suggested by the research about leadership in baby-faced and gay Black men. Additionally, the prototypical leader is a White man, where White men are rated as more effective leaders compared to individuals within other racial groups (Rosette, Leonardelli, & Phillips, 2008), so Black men must be perceived as especially trustworthy and competent to overcome these biases. Based upon these premises, perceptions of dominance and trustworthiness from the voice in combination with perceived threat potential based upon stereotypes about Black people will likely affect ratings of their leadership ability. Specifically, Black men with vocal characteristics that elicit trust and decrease perceptions of dominance may be attributed leadership traits to a greater extent than Black men with voices that are perceived as threatening. Although previous research suggests that voice pitch has an effect upon leadership selection, where male CEO's with lower-pitched voices tend to be more successful (Mayew, Parsons, & Venkatachalam, 2013), there is no research examining the interaction between race and voice pitch upon perceived leadership ability.

The current study makes an important contribution by examining the effects of stereotypes and vocal characteristics upon one's success in leadership positions and their perceived threateningness. We focus on vocal characteristics because the voice can be modulated volitionally (Fraccaro et al., 2013; Hughes, Mogilski, & Harrison, 2014; Pisanski et al., 2016), and individuals are constantly provided auditory feedback during speech, which facilitates precision in encoding. This is in stark contrast to encoding of facial expressions, which cannot be monitored without outside assistance, so it is more difficult to exert as much control over encoding intended facial expressions effectively. Through vocal modulation, individuals can exert precise control over how others perceive them, which may facilitate their social goals (Fraccaro et al., 2011). In the case of Black men, they may

modulate their voice to reduce perceptions of threat and increase perceived leadership ability in settings where they are the minority, like corporate America. If this is the case, it is imperative to determine whether vocal characteristics can serve as a disarming mechanism, which underlies the goals of the current experiment.

Through our experiment, we examined whether voice pitch differentially modulates perceptions of threat and leadership ability for Black and White men by creating recordings for participants, then randomly assigning them to four conditions with different voices and names, which served as our manipulation of group membership and voice pitch, respectively. We chose to exclude women from the sample of stimuli for this study because we anticipated that the interaction effect between race and voice pitch would be stronger amongst men. According to the out-group male target hypothesis (Navarrete, McDonald, Molina, & Sidanius, 2010), out-group men are more likely to be perceived as threatening since men were more likely to engage in inter-group conflict throughout our evolutionary history. Empirical evidence shows that Black men are more likely to be perceived as a threat to physical safety compared to Black women (Sidanius & Veniegas, 2000), which can amplify fearful responding towards Black men when an individual feels vulnerable to threat (Maner et al., 2005).

We hypothesized that participants would rate Black men with high-pitched voices lower on traits associated with threat and higher on traits associated with leadership ability compared to Black men with low-pitched voices. On the other hand, White men with a low-pitched voice will be rated higher on traits associated with leadership ability compared to White men with a high-pitched voice, as suggested by previous research (Klofstad, Anderson, & Peters, 2012). For our secondary hypotheses, we anticipated that perceived trustworthiness would be negatively related to perceived threat, while perceived dominance would be positively related to perceived threat.We also expected main effects of race and voice pitch upon perceived trustworthiness and perceived dominance, where Black men would be perceived as less trustworthy, while low-pitched voices would be rated as

more dominant.

Methods

Participants

527 participants originally completed the survey, and we excluded any participants that alluded to the hypothesis in the suspicion check by responding that the study was about race and/or race and voice perceptions (N = 20; 3.8% of total sample). The final sample consisted of 507 (278 Women, 229 Men) participants from Amazon Mechanical Turk (see Table A1 for demographic information). Ages ranged between 19 and 82 years (_M___age_= 40.07, SD = 13.26). We included participants based upon the following criteria: (a) adults on (b) Amazon Mechanical Turk) (c) born and currently residing in the US (d) have had 90% or greater of their previous HITs approved, and (e) have a device with audio capabilities. We excluded Black individuals during the pre-screening process, since we are primarily interested in understanding the factors that affect threat and leadership perceptions of Black men, and group membership may differentially affect these perceptions.

Design

The study was a 2X2 within-subjects design with two independent variables: voice pitch (high or low) and race (White or Black names). Each of the four conditions was counterbalanced. Names and individual voices were randomly assigned to participants without repeat. This ensured that individuals would not listen to a high and low voice that resulted from the same original voice.

Procedure

Participants were recruited from Amazon Mechanical Turk by posting a HIT (human intelligence task) on the site. They were told that they would listen to a participant that previously provided their recording and took a "series of character trait and performance tests," which would then be compared to the participants' ratings to assess the accuracy of their perceptions. Upon being assigned to a recording, they learned the participant's name, and were provided other information about the recording (i.e., location, date) to make the design less conspicuous. Then, they listened to the participants' recording by clicking on the Soundcloud file embedded in the survey.

All of the names were randomly assigned to correspond to the high-pitched or the low-pitched conditions. The presentation of the four names for the recordings was randomized and counterbalanced across participants. We verified that the randomization worked by comparing the number of participants that were assigned to each condition, which were relatively uniform. The four conditions (Black name high pitch, Black name low pitch, White name high pitch, White name low pitch), were equally presented first, second, third, and fourth (see Table A2).

They were asked to assess the participant's character based upon their voice using a series of 100-point slider scale questions (i.e., trustworthiness, dominance, threateningness), which served as our measures of perceived trustworthiness, perceived dominance, and perceived threat, respectively. The presentation of the scale items was counterbalanced for each participant and within each condition. Additionally, we asked them to rate the individuals in the recording on various traits that were independently rated as important for leaders on 100-point slider scale items. Finally, they indicated their preferences for engaging in different types of interdependent relationships with the people in the recording on 100-point slider scales. Participants could listen to the recordings as frequently as they desired before rating the voices. They completed demographic questions and indicated

what they thought the study was about as a suspicion check. After participants completed the suspicion check, we determined whether the manipulation of the names elicited perceptions of the race of the recorded individuals through a series of manipulation check questions. First, we created a name attention check score based upon whether the participants remembered the names of the people in the recordings. The participants were presented with a list of eight names, four of which were included in the study. Every time they correctly identified a name that was presented to them during the study, they received a point, for a total name recall score of four points ($M = 3.09$, $SD = .979$). If they incorrectly selected a name that was not presented to them, they did not receive a point. On the name attention check, participants recalled the Black names with greater accuracy (68.34%) than White names (59.69%). Also, they remembered the conditions presented first (69.09%), second (70.80%), and fourth (76.30%) better than they remembered the condition presented third (39.88%). Participants were asked how many people in the recordings they thought were White or Black (see Figures A1 and A2). Finally, we asked participants to rate the likelihood that people with the names used in the study would be White or Black (see Table A3). A debriefing page explaining the true purposes of the study and the logic behind the deception was provided before payment. Participants were paid $1.00 for their participation.

Materials and measures

Voice stimuli. For the voice stimuli, we recorded the voices of eight White men between 18-30 years of age in Audacity using the Zoom H4N Handy Recorder with a sampling rate of 44.1kHz. The men quoted the first sentence of the Rainbow Passage (e.g., "When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow") (Fairbanks, 1960). At the end of each sentence, the men read a randomly assigned identification number provided by the researchers. The four-digit identification numbers were created randomly, and participants were required to enter the identification number

as a means of verifying that they were listening to the recordings.

After the recording sessions, each voice was manipulated to have a higher or lower pitch in Praat (Version 6.0.36) (Boersma & Heuven, 2001), which served as our manipulation of threat potential through the voice. We followed the standard methods in voice research by raising and lowering each voice by 0.5 equivalent rectangular bandwidths (ERBs) using the Pitch-Synchronous Overlap Add tool in Praat, which produces a shift in perceived pitch of approximately 20 Hz in either direction (e.g., Apicella & Feinberg, 2009; Klofstad et al., 2012; Tigue, Borak, O'Connor, Schandl, & Feinberg, 2012; Vukovic et al., 2010). We set the pitch floor to 70 Hz and the pitch ceiling to 250 Hz, which has been validated as an appropriate range for male voices (Vogel, Maruff, Snyder, & Mundt, 2009). Many researchers manipulate ERB instead of Hertz because a change in pitch is perceived differently depending upon the original pitch that was manipulated, since there is a logarithmic relationship between actual pitch and perceived pitch (Stevens, 1998). Also, the ERB manipulations will not affect other acoustic characteristics of the recording (e.g., rate, intensity) (Feinberg, Jones, Little, Burt, & Perrett, 2005). Since each of the voices was raised and lowered in pitch, there were a total of sixteen manipulated recordings. We checked the manipulation by comparing the mean pitch for the original voices (M = 104.37, SD = 14.09) to the lower manipulations (M = 90.21, SD = 9.79) and the higher manipulations (M = 121.30, SD = 17.28). All of the manipulated files were uploaded to separate Soundcloud links and embedded in the survey.

Names for race manipulation. To manipulate perceptions of race, we used four names that are typically associated with Black people (i.e., Tyrone, Keyshawn, Deshawn, Terrell) and four names that are typically associated with White people (i.e., Scott, Brad, Brett, and Logan) (Gaddis, 2017). Names were presented before the participants listened to the voice recording. Each name was chosen based upon the criteria that 90% or more of raters from Gaddis (2017) thought that the individual was either Black or White when they were asked about their perceptions of the person's race based upon their name.

Perceived leadership ability. We recruited 55 participants on Amazon Mechanical Turk to serve as independent raters for identifying the leadership traits used in the experiment. We provided them with a list of fifteen traits from which they could select what they considered most valuable for successful leaders of businesses and companies (e.g., drive, creativity, confidence) (Kirkpatrick & Locke, 1991). We selected the traits that were ranked, on average, in the 30th percentile of responses (where 1 is considered the most important trait for a leader). The traits that were selected for the leadership composite score based upon these criteria were intelligence, effective communication, confidence, and problem-solving ability, which were rated by participants in the final study using 100-point slider scale items.

To create the leadership ability composite, we averaged participants' ratings of the individual in the recording on the four traits. Higher scores denote greater perceived leadership ability. The measure had high internal consistency across participants in the final sample (alpha = .91; averaged across all conditions).

Perceived threat, trustworthiness, and dominance. Single questions were used to elicit perceived threat, trustworthiness, and dominance. Participants responded using a 100-point slider scale.

Multilevel models

Given the nature of the data (condition nested within subjects and subjects nested within the "names" associated with each voice), we employed multilevel models (also known as linear mixed-effects models or hierarchical linear models) (**???**; Finch, Bolin, & Kelley, 2014; Garson, 2013; Gelman & Hill, 2007; Raudenbush & Byrk, 2002) to analyze the data, when feasible. The basic premise of using this type of analysis is to account for the inherent correlation among the observations nested within other variables. For instance, within the context of the current study, we measured participants' rating of threat across all conditions, so it is likely perceptions of threat within each participant will

be correlated, since there may be inherent individual differences in participants' baseline perceptions of others' threat and/or perceptions of threat in response to each combination of race and voice pitch. If we did not account for this correlation in responses within each participant, we would be violating the assumption of independence of observations. Repeated measures analysis of variance (ANOVA) is often used for analyzing data of this nature, however we use multilevel models because they present several notable advantages over repeated measures ANOVA. For instance, multilevel models are more powerful in the face of "unbalanced" repeats, where the measure of interest is missing one or more observations (within reason). Instead of employing listwise deletion in the face of missing data points during analysis, like repeated measure ANOVA, which reduces the effective sample size, multilevel models use the data available within a group to estimate parameters and compute inferential statistics, while accounting for the fact that some estimates are more reliable than others (Brauer & Curtin, 2018; Misangyi, LePine, Algina, & Geoddeke Jr., 2006; Raudenbush & Byrk, 2002). Notably, we assume that the data points are missing at random for these inferences. Additionally, multilevel models allow researchers to explore multiple groups with correlated observations (that is, multiple sources of nonindependence) (Brauer & Curtin, 2018; Westfall, Judd, & Kenny, 2015; Westfall, Kenny, & Judd, 2014), while repeated measures ANOVAs only allow one to account for one source of nonindependence (e.g., Baayen, Davidson, & Bates, 2008; Judd, Westfall, & Kenny, 2017). The effects of continuous predictors that may be clustered can only be analyzed using multilevel models, since repeated measures ANOVAs only accept within-group (e.g., subject) factors as predictors (Brauer & Curtin, 2018; Misangyi et al., 2006). Finally, multilevel models allow researchers to explicitly model different sources of variation within the data (e.g., individual and group-level variation in group-level estimates, variation in individual-level estimates) and estimate the effects for specific groups (Gelman & Hill, 2007).

Model estimation and comparison. Multilevel models are a variation of classical regression that assign a probability model to specific regression coefficients (Gelman & Hill, 2007). The parameters of the second-level model have their own coefficients, known as hyperparameters (Gelman & Hill, 2007). Although classical regression has the capacity to model varying coefficients with the use of indicator variables, multilevel models are unique in their ability to model the variation between groups by including varying coefficients and models for each varying coefficient (Gelman & Hill, 2007). To model variation at multiple levels, these models incorporate what are known as "fixed" and "random" effects, where random effects are typically conceptualized as effects that vary across the nested groups, while fixed effects are constant across all groups (Finch et al., 2014).

When analyzing data using multilevel modeling, there are different random effects structures that can be used to model the data. A random effects structure is essentially the way the parameters are assumed to vary across the nested groups (Barr, Levy, Scheepers, & Tily, 2013). The most basic random effects structure includes only a random intercept, which essentially allows the intercept to vary across groups (i.e., there is a different intercept estiamted for each group). A more complex random effects structure involves allowing the slopes (i.e., fitting a unique regression line to each group) and intercepts to vary by group. There are many different ways to model the random effects structure (e.g., random slope by group with correlated intercepts, random slope by group without varying intercepts by group, uncorrelated random intercept and slopes by group, etc.) (Meteyard & Davies, 2020), which will change the interpretation of the results and even reduce power or increase Type 1 error (Barr et al., 2013; Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017), so it is important to identify a random effects structure that is appropriate for the data.

The literature on determining random effects structure for one's data is mixed. Some have argued that it is imperative to fit maximal models (i.e., fit random slopes and intercepts for each predictor in the model, including interactions) whenever possible, and

only reduce the random effects structure when the model does not converge (Barr et al., 2013). When a model does not converge, it essentially means that the optimization algorithm used to estimate parameters cannot reliably determine the maximum likelihood function for the current model (Brauer & Curtin, 2018). This typically occurs when there is insufficient data for the number of parameters being estimated. Thus, failures to converge are much more likely to occur when trying to estimate maximal models. On the other hand, fitting models with only intercepts varying by group leads to inflated Type 1 error (Barr et al., 2013; Schielzeth & Forstmeier, 2009). Thus, Barr et al. (2013) argue that the common practice of fitting varying-intercept only models can lead to biased conclusions and instead argued for maximal models. However, it has been argued that the rise in employing "maximal" models can lead to their own set of problems, namely i) failure to converge (Bates, Kliegl, Vasishth, & Baayen, 2015), ii) models that converge but are so overparameterized that they are uninterpretable (Bates et al., 2015), and iii) loss of power due to random effects contributing little to the model (Matuschek et al., 2017). In place of "maximal" models, Matuschek et al. (2017) and Bates et al. (2015) argue for "parsimonious mixed models," where the researcher uses a pre-determined model comparison technique (e.g., likelihood ratio test, Akaike information criterion, Bayes/Schwarz information criterion) to select the random effects structure that best fits the data. For selecting parsimonious mixed models, one would first fit a maximal model, then remove random effects that are not contributing to the model (i.e., variance is close to 0), stopping before they reach a model that would significantly reduce the goodness of fit (Bates et al., 2015; Matuschek et al., 2017). In support of this argument, Matuschek et al. (2017) use simulations to demonstrate that parsimonious models can reduce Type 1 error associated with underfitting models, while attaining higher power than maximal models. This is because maximal models can lead to a decrease in statistical power if they have random effects with variances near 0 which do not contribute to the fit of the model but reduce the degrees of freedom, essentially increasing the standard errors of the fixed effects

estimates (Matuschek et al., 2017). At the same time, it is generally accepted that random effects with near-zero variance do not affect goodness of fit tests (Brauer & Curtin, 2018).

Despite the lack of consensus regarding how to determine the final model, most researchers suggest starting with the maximal model and that the final model should have an effects structure that aligns with the researcher's main hypothesis, even if these random effects have near-zero variance (Barr et al., 2013; Bates et al., 2015; Brauer & Curtin, 2018).

Another general point of consensus is that when fitting models with varying effects, one should use restricted estimated likelihood (REML) instead of maximum likelihood (ML) estimation for unbiased estimates of the random effects parameters (Brauer & Curtin, 2018; Browne & Draper, 2006; Elff, Heisig, Schaeffer, & Shikano, 2020; Gelman & Hill, 2007; Maas & Hox, 2005), especially with smaller samples at the group-level (**???**; **???**; McNeish, 2017). The problem with ML estimation typically arises with smaller samples because the process of ML estimation tends to ignore variability in the fixed effect estimates and does not account for the degrees of freedom (DF) used to estimate the fixed effects (McNeish, 2017). These effects of using ML can lead to more bias in random effects estimation with smaller samples because they are more sensitive to small changes in the degrees of freedom and tend to have larger sampling variability. Since random effects parameters are estimated based on the fixed effects parameters, this can cause them to be underestimated (McNeish, 2017). As a result, the standard errors of the fixed effects tend to be underestimated because the random effects estimates are integrated into the formula for fixed effects standard errors. With smaller standard errors, the t or Z test statistic will be overestimated, leading to higher Type 1 error. The process employed by REML leads to better estimates of the random effects, which in turn improves the fixed effects standard error estimates (McNeish, 2017).

Finally, researchers have examined how different techniques for evaluating significance

of multilevel models affect Type 1 error rates. Notably, Luke (2017) show through simulations that likelihood ratio tests and applying the Z distribution to the Wald t values from the model output can lead to unusually high Type 1 error rates, especially with smaller samples (i.e., less than 40-50 number of items and/or subjects). This is the case when fitting models using both ML and REML. Of the options available to researchers in R, Type 1 error is closest to .05 when deriving p-values using Kenward-Roger (**???**) or Satterthwaite (Satterthwaite, 1941) corrections for approximating for denominator degrees of freedom for F statistics or DF for t statistics (Luke, 2017). Although these corrections tend to produce similar output (Luke, 2017), (**???**) and McNeish (2017) argue that the Kenward-Roger provides slightly better approximations by correcting standard error and estimating DF, while the Satterthwaite correction only estimates the effective DF.

Sample size considerations. Another important consideration in determining model structure is the sample size at each level (i.e., number of groups and number of individuals within each group) of multilevel models. Like in most parametric statistical inference, the estimates become unreliable (and in some cases, impossible to estimate) with sparse data. These effects may differ depending on the level, where Scherbaum and Ferreter (2009) showed that increasing level-2 (i.e., number of groups) sample size had a larger effect on variance components than increasing level-1 (i.e., number of individuals within each group) sample size.

Understanding how "sparse" one's data can be at each level while being able to maintain unbiased estimates has been the subject of several lines of recent work. One of the seminal pieces in this literature suggested that level-2 standard errors are biased when the sample size is less than 50 (i.e., there are fewer than 50 groups) (Maas & Hox, 2005). Other recommended standards are to have 10 observations for at least 100 groups to estimate a random intercept for said group, and at least 20 observations with a minimum of 200 groups for estimating slope variance (Clarke & Wheaton, 2007). Scherbaum and Ferreter (2009) recommends 30-50 trials/items per participant for power. Notably, the size

of each cluster can affect estimation of the random effects, but tend to have little to no impact on estimation of fixed effects (Clarke & Wheaton, 2007; Maas & Hox, 2004, 2005).

Thus, recent work has focused primarily on the effects of small samples on the estimation of random effects. As of recently, many researchers are suggesting that, under certain conditions (e.g., continuous outcome variables, five or fewer fixed effects, no missing data, one or two variance components), there can be as few as 7-10 groups at the second level to be able to estimate random effects with reasonable accuracy using REML with cross-sectional data. However, the appropriate sample size will intrinsically depend on the nature of the data and model at hand (**???**).

With small samples, some suggest that Bayesian estimation can be a more accurate alternative (Stegmueller, 2013) because Bayesian statistics do not rely on the central limit theorem (**???**). Specifically, Stegmueller (2013) performed a Monte Carlo experiment to compare the performance of frequentist and Bayesian multilevel models when there are few (e.g., 5, 10, 15) groups, and showed that the frequentist approach tended to be anti-conservative and biased with smaller samples. However, in a response to Stegmueller (2013), Elff et al. (2020) has recently argued against the notion that standard multilevel models are inferior to models following a Bayesian framework with a small number of groups. Specifically, they showed that the estimation bias found in Stegmueller (2013) could be solved by simply using i) REML estimators for variance parameters and ii) a t-distribution with appropriate degrees of freedom for statistical inference. Through these steps, the standard multilevel models were found to produce unbiased estimates.

Additionally, any possible advantages of Bayesian estimation are completely dependent on the choice in a prior probability distribution (often simply called a prior) (**???**; Gelman & Hill, 2007). A prior essentially represents the knowledge, a priori, one has (if any) about the distribution of the parameters, which are then combined with the data observed to produce posterior inferences. Thus, the choice of a prior probability

distribution is critical, especially with smaller samples (McNeish & Stapleton, 2016). For instance, in a systematic review of the literature on estimation while using small samples, Smid, Mcneish, Miočević, and Schoot (2020) showed that inference with uninformative priors can lead to estimates that are just as, if not more, biased, than the estimates from frequentist methods when working with a small number of groups. Unfortunately, the default prior distribution in most software is uninformative, so it is entirely possible that many researchers acquire biased estimates by naively using Bayesian estimation (**???**).

LMEM methodology for current research. Then, describe how I estimated each model, given the current literature. We used R (Version 3.6.3; **???**) for analysis. To determine the maximal effects structure, we follow the guidance of Brauer and Curtin (2018) and Barr et al. (2013), who argue that i) every source of nonindependence should be modeled through a random intercept ii) generally, there should be a random slope for each within-unit predictor and, iii) one should estimate random slopes for interaction effects when all factors comprising the interaction are within-group. They also note that there are exceptions to these general rules of thumb. For instance, i) does not need to be followed if the purported random effect is fully confounded with a predictor in the model. That is, we would exclude groups where the random variable is nested within a fixed effect. Since the names chosen for the current experiment were necessarily nested within race, we do not model random effects for name when race is included as a fixed effect.

See Appendix Table X for complete report of all models compared.

We tested the assumptions of each model (i.e., linearity, random distribution of residuals, homoscedasticity, absence of collinearity).

Assumptions specific to multilevel models: * level 2 residuals are independent between clusters (aka random intercept & slopes at level 2 are independent of one another across clusters). so one subjects' slope & intercept should not be strongly related to other subjects' slope & intercept * Level 2 intercepts & coefficients are assumed to be

independent of the level 1 residuals (ie errors for cluster level estimates are unrelated to errors at the individual level). aka how much you are wrong in predicting one's specific response from subject 5 should not be related to how much you are wrong in predicting subject 5's overall slope & intercept * level 1 residuals are normally distributed & have constant variances * level 2 intercepts & slopes have a multivariate normal distribution with a constant covariance matrix

(Finch et al., 2014)

Results

Effects on perceived threat

report what data cleaning has been completed, outlier/data removal, transformations. report sample size entered in terms of total number of data points & various sampling units (ie subjects, number of groups specified as random effects). report wether models meet assumptions for LMMs

To determine the effect of voice and race upon perceptions of threat, we ran a 2 (voice pitch: high or low) X 2 (race: Black or White) repeated measures ANOVA with perceived threat as the dependent measure (see Figure 1). There was a main effect of voice pitch upon perceived threat, F (1, 506) = 62.225, p < .001, eta2_p_= .11. Race did not significantly predict perceived threat, F (1, 506) = 0.170, p = .680, eta2_p_= .000, and the interaction between the variables was not significant, F (1, 506) =1.351, p = .246, eta2_p_= .003. We ran post-hoc tests with Bonferroni corrections to explore the main effect of voice pitch, which showed that low voices (17.87 ± .77) were perceived as significantly more threatening compared to high voices (12.34 ± .61), with a mean difference of 5.49(95% CI, 4.13 to 6.86).

Figure 1. Mean perceptions of threat as a function of voice pitch and perceived race. Error bars represent 95% confidence intervals. The perceptions of threat items were on

100-point scales.

Effects on perceived leadership ability

To determine the effect of voice and race upon perceived leadership ability, we ran a 2 (voice pitch: high or low) X 2 (race: Black or White) repeated measures ANOVA with the leadership composite score as the dependent measure. Both voice pitch, F (1, 506) = 10.109, p = .002, eta2_p_= .02, and race, F (1, 506) = 10.622, p = .001, eta2_p_= .021, significantly predicted leadership composite ratings (see Figure 2). Black voices (61.72 ± .73) were rated higher on leadership traits compared to White voices (59.12 ± .72), with a mean difference of 2.57 (95% CI, 1.03 to 4.16). On the other hand, lower voices (61.72 ± .74) were rated higher on leadership qualities than higher voices (59.12 ± .72), with a mean difference of 2.60(95% CI, .99 to 4.21). The interaction was not significant, F (1, 506) = 0.079, p = .779, eta2_p_= .000. See Table A4 for mean differences in perceived leadership ability between White and Black voices based upon participants' demographic characteristics.

Figure 2. Mean perceptions of leadership traits as a function of voice pitch and perceived race. Error bars represent 95% confidence intervals. The perceptions of leadership items were on 100-point scales.

Trustworthiness and dominance predicting threat

The relationship between trustworthiness and dominance with perceived threat was examined by running a multiple linear regression with the averaged ratings across conditions. Overall, the predictors explained 19.7% of the variance in perceived threat, F (2, 504) = 61.86, p < .001, adjusted _R___2_ = .194. We found that trustworthiness was negatively related to threat, b = -.28, t (504) = -6.72, p < .001, while dominance was positively related to threat, b = .44, t (504) = 10.46, p < .001. We also ran the

regression across all four conditions, confirming that the effect was present in each cell of our design (see Table A5).

Effects on trustworthiness and dominance

To determine the main effect of race upon perceived trustworthiness while controlling for the effects of voice pitch, we ran a 2 (voice pitch: high or low) X 2 (race: Black or White) repeated measures ANOVA with perceived trustworthiness as the dependent measure to examine whether race altered perceptions of trustworthiness. There was a significant main effect of race, F (1, 506) = 7.04, p = .008, eta2_p_= .01, upon perceived trustworthiness, where Black men (61.93 ± .70) were perceived as more trustworthy than White men (59.80 ± .79), while controlling for voice pitch and the interaction term. The mean difference in ratings was 2.13 (95% CI, 0.55 to 3.70).

We also tested whether voice pitch predicted perceived dominance independent of the effects of race by running a 2 (voice pitch: high or low) X 2 (race: Black or White) repeated measures ANOVA with perceived dominance as the dependent measure. Although the effect of voice pitch on perceived dominance was non-significant, F (1, 506) = 3.49, p = .062, eta2_p_= .007, there was a significant main effect of race upon perceived dominance, F (1, 506) = 68.25, p < .001, eta2_p_= .12. White men were perceived as more dominant (46.15 ± .94) compared to Black men (37.42 ± .84), with a mean difference of 8.73 (95% CI, 6.65 to 10.81).

Although some of the above tests failed assumption checks because they had outliers, we re-ran the tests without them and found the same effects. Therefore, we reported the results of the original tests with outliers included.

Exploratory analyses

Effects on individual leadership traits. We control for multiple hypothesis testing in all exploratory analyses by setting our criteria of significance at the .01-level. First, we ran a series of 2 (voice pitch: high or low) X 2 (race: Black or White) repeated measures ANOVAs with each of the leadership traits as dependent measures to break down the leadership composite effects. Participants were more likely to perceive Black recordings (62.07 ± .84) as effective-communicators compared to White recordings (58.64 ± .82), F (1, 506) = 12.66, p < .001, eta2_p_= .024, with a mean difference of 3.43 (95% CI, 1.54 to 5.33). Also, participants were more likely to perceive low-pitched recordings (58.02 ± .82) as problem-solvers compared to high-pitched recordings (55.60 ± .79), F (1, 506) = 7.63, p = .006, eta2_p_= .02, with an average difference of 2.42 (95% CI, 0.70 to 4.14).

There was a significant effect of voice pitch upon perceived confidence, F (1, 506) = 20.94, p < .001, eta2_p_= .04, where low voices were perceived as significantly more than confident (61.64 ± .87) than high voices (56.74 ± .83), with an average difference of 4.90 points for the two groups (95% CI, 2.80 to 7.00). Also, race had a significant effect on perceived confidence, F (1, 506) = 11.32, p = .001, eta2_p_= .022, where Black men were perceived as significantly more confident (60.89 ± .81) than White men (57.48 ± .84), with ratings differing by 3.41 points on average (95% CI, 1.42 to 5.39).

Main effect of race on perceived leadership ability. Given the unanticipated findings that Black men were rated as better leaders, we explore the data to examine three possible explanations for our outcomes: social demand effects, contrast effects, and the effects of stereotypes about Black men (i.e., more dominant and aggressive). To examine these possibilities for their validity in explaining our pattern of results, we will describe the evidence for and against each explanation based upon our exploratory analyses.

Social demand effects.

First, social demand effects are always a concern when running within-subjects

studies about race, because people are averse to being considered biased against Black people. When participants were presented 2 White names and 2 Black names (in a random order), it is entirely possible that they guessed that the study was focused on perceptions based upon race. Although we included a suspicion check and excluded participants based upon stringent criteria, the suspicion check may have biased participants to indicate that they were not suspicious. Specifically, they had to type in a text entry box if they were suspicious about the hypotheses, whereas they could simply select a multiple-choice option to indicate that they were not aware of the hypotheses. As a result, participants may have chosen the easier multiple-choice option on the suspicion check instead of choosing to type in their actual prediction of the purpose of the study.

To explore the plausibility of social demand effects as a potential explanation for our hypotheses, we tested a series of assumptions that we assumed would hold if participants were responding in a socially desirable way. If social demand effects were underlying our results, we would expect participants to rate the Black voices higher on leadership if they remembered the Black names better (i.e., performed better on the manipulation check). To test this assumption, we compared effects of the number of Black names that participants remembered (0, 1, or 2) upon the perceptions of leadership for the Black voices averaged across conditions by running a one-way ANOVA. This test suggested that performance on the manipulation check (i.e., memory for the Black names) did not significantly affect perceived leadership, F (2, 504) = 1.00, p = .37, eta2_p_=.004.

The study also included interdependent relationship measures for exploratory analyses, where participants were asked how much they would like to engage in different types of interdependent relationships (i.e., work project team member, close friend, neighbor, employee) with the person in the recording using 100-point slider scale items. For these measures, we would also expect to see higher ratings of Black voices in relationships where Black people tend to be disadvantaged (e.g., employee or work project team member) because participants would try to avoid appearing biased. We ran a two-way

(race by voice pitch) ANOVA with their preferences for having the recorded individuals as employees or work project team members. There was no significant effect of race on preferences for employees, F (1, 506) = 3.36, p = .07, eta2_p_= .007, or work project team members, F (1, 506) = 1.94, p = .16, eta2_p_= .004. This contradicts what we would expect for participants that are responding in any socially desirable way.

Other effects we would expect if participants were engaging in socially desirable responding are higher ratings for Black voices on trustworthiness and lower ratings on dominance and threat. Specifically, the prominent stereotype that Black men are criminals would prompt participants to rate them higher on trustworthiness if they did not want to appear biased. Along similar lines, perceptions of threat and dominance are a major stereotype that are applied to Black men (Quillian & Pager, 2001), suggesting that participants should rate Black men lower on these traits to avoid appearing biased, if they correctly guessed our hypothesis. We found a significant effect of race upon perceived trustworthiness, F (1, 506) = 7.04, p = .008, eta2_p_= .014, where Black men (61.93 ± .70) were rated higher on perceived trustworthiness compared to White men (59.80 ± .80), with a mean difference of 2.13 (95% CI, 0.55 to 3.70). Also, there were significant differences in ratings for perceived dominance between the races, F (1, 506) = 68.25, p < .001, eta2_p_= .119, such that White men (46.15 ± .94) were rated higher on perceived dominance compared to Black men (37.42 ± .84), with a mean difference of 8.73 (95% CI, 6.65 to 10.81). These results provide support for socially desirable responding. However, contrary to what we would expect for participants that were trying to avoid responding in a biased manner, there was no significant effect of race on threat (see Figure 1 above). In sum, there is both evidence in favor of and against social demand effects in explaining the unexpected effect of race on perceived leadership.

Contrast effects.

Another potential explanation for our unexpected results is contrast effects, which are based upon the shifting stereotypes model (Biernat, Manis, & Nelson, 1991). This model

posits that an individual will judge others on stereotype-relevant dimensions relative to other individuals within their social category. In the case of our study, the order of presentation of the name and voice stimuli may have affected the outcomes, since the stereotypical Black names were presented before the voices. This may have preempted them to expect a voice that sounded relatively uneducated. Previous research shows that the voice can convey SES and education levels (Kreiman & Sidtis, 2011), so it is entirely possible that participants used this information in their assessments of the individuals in the recordings. Since the individuals that we recruited for our voice stimuli were generally well-educated (University of Pennsylvania graduate students and upper-level undergraduates) relative to the general population, the Black voices might have exceeded their low expectations, eliciting higher ratings. On the contrary, White men tend to have positive stereotypes attributed to them regarding their leadership ability (Rosette et al., 2008), so the baseline expectations for the White voices were relatively high, which may have also contributed to the effect of race upon perceived leadership ability.

If contrast effects can explain our results, we would expect to find a similar effect for Black men on the threat measure, where they would be rated significantly lower on perceived threat because their voice undermines the stereotype that Black men are extremely threatening. However, there is no main effect of race on perceived threat (see Figure 1). Since we created the leadership composite, it is entirely possible that it was not a valid representation of the traits that are integral to a leader. However, our leadership composite is strongly correlated with the boss measure, r (505) = .66, p < .001, suggesting that it is a valid measure of leadership ability. Furthermore, we would expect participants to rate Black men higher on the boss measures if contrast effects contributed to our results. Contrary to this possibility, we find that participants did not exhibit any differential preferences based upon race, F (1, 506) =3.35, p < .068, eta2_p_= .007. Additionally, as the average ratings for perceived leadership increase, there should be a greater discrepancy between Black and White ratings, since we would expect contrast

effects to be greater as the voices deviate more from what an individual expects based upon their stereotypes. We found this effect for almost all voices except for the voice that was rated highest on perceived leadership (see Figures A3 and A4).

Effects of stereotypes about dominance and aggressiveness.

The final explanation that we explored in our data was the possibility that Black men were rated higher on leadership because, in the absence of threat, they may benefit from stereotypes that typically attribute dominance and aggressiveness to their social group (Devine & Elliot, 1995). Dominance and aggressiveness were invaluable characteristics for leaders throughout our evolutionary history who needed to be successful during lethal inter-group conflicts and dangerous hunting sessions (Van Vugt et al., 2008a). Therefore, any personal characteristics that are perceived as more dominant or aggressive may increase perceived leadership ability, even though these traits may not accurately reflect leadership ability in the modern day (Klofstad & Anderson, 2018; Li, Vugt, & Colarelli, 2017). In this way, Black men might have been rated as better leaders because they were attributed dominance and aggressiveness to a greater degree than White men based upon stereotypes. If this explanation is valid, we would expect Black men to be rated higher on dominance, which, as demonstrated above, is not what we find in our data. Additionally, we would expect that participants would prefer Black men as a boss. Our analyses above do not support this assumption.

Discussion

Overall, we found that voice pitch has a significant effect upon perceived threat and leadership, where lower-pitched voices were rated as more threatening and better leaders compared to their higher-pitched counterparts, replicating previous literature on this topic (Hodges-Simeon et al., 2014; Puts et al., 2012). We also found an unexpected effect of race upon leadership, where Black men were rated significantly higher on perceived leadership compared to White men. This finding contradicted our expectations, since most other

research on this topic suggests that White men are prototypical leaders and are much more likely to be rated higher on perceived leadership than other social groups (Rosette et al., 2008). Our original primary hypotheses were not supported, since we did not find the expected interaction effects of voice pitch and race on perceived threat or perceived leadership. With regards to our secondary hypotheses, we find support for our prediction that perceived trustworthiness and perceived dominance would be related to perceived threat in the expected directions. Specifically, perceived trustworthiness was negatively related to perceived threat and perceived dominance was positively related to perceived threat, even when we examined the relationship broken down by each condition, suggesting that it is a robust effect. This aligns with previous research examining the effect of the facial dominance and trustworthiness that combine to affect perceived threat (Oosterhof & Todorov, 2008), but no studies have examined this relationship based upon vocal characteristics and race before. Therefore, our study provides preliminary support for the notion that the observed effects of facial trustworthiness and dominance on threat can be generalized to other personal characteristics (i.e., the voice).

Regarding our other secondary hypotheses, we found a significant effect of race upon perceived trustworthiness, but in the opposite direction of our expectations, where Black men were rated as significantly more trustworthy compared to White men. Most of the literature in this domain suggests that Black men are perceived as less trustworthy (Stanley, Sokol-Hessner, Banaji, & Phelps, 2011; Stanley et al., 2012), largely because of negative stereotypes that are applied to their social category. On the contrary, we did not find the expected effect of voice pitch upon perceived dominance, but instead found an unexpected effect of race, where White men were rated as significantly more dominant compared to Black men.

Since our findings regarding race and leadership ability were especially unexpected given previous research, we explored three potential explanations for the results. First, we showed that there is mixed evidence in favor of social demand effects upon our findings,

where we did not find that participants rated the Black voices higher on leadership if they remembered the Black names better (i.e., performed better on the manipulation check), nor did they rate Black men higher for specific leadership positions (i.e., as a boss). These are assumptions we would expect to hold if participants were engaging in socially desirable responding. On the other hand, they rated Black men higher on perceived trustworthiness and lower on perceived dominance, which aligns with what we would expect if they were trying to avoid appearing biased against Black men. Although these analyses provide some support for socially desirable responding, participants did not rate Black men lower on perceived threat, which is a prominent stereotype applied to this social category that participants may try to avoid confirming if they were in fact concerned about being labeled as biased.

We also explored the possibility that contrast effects could explain our findings, where the order of presentation of the voice and name stimuli (i.e., names presented before voices) may have affected participants' ratings. Specifically, the shifting stereotypes model (Biernat et al., 1991) posits that individuals within a negatively stereotyped social category will be rated higher on subjective ratings for stereotype-relevant traits because they are judged relative to others within their respective social category and as a result, have a lower threshold to surpass. In the case of our study, the names may have activated expectations about the vocal characteristics that they would hear in the recording, and when the voices completely exceeded those low expectations, they were rated as subjectively "superior" to their White counterparts who had a higher threshold to exceed. Based upon these premises, we would expect that there should be a similar contrast effect for Black men on the threat measure, but there is no main effect of race. Additionally, voices that are rated higher on leadership should show a greater discrepancy between Black and White ratings, which we found across most voices, except for the voice that was rated highest on leadership. In that case, the voice was rated higher when it was assigned a White name compared to a Black name. It is possible that this voice deviated from the

general pattern either because the pattern would not exist if we had sampled from a larger group of voices or because the voice was unique in its characteristics.

The third potential explanation that we explored was that the stereotypes that Black men are aggressive and dominant may have conferred higher ratings for the recorded individuals on perceived leadership, but only in the absence of threat. When individuals have personal characteristics that are perceived as dominant and aggressive, they are more likely to be selected as leaders, which is based upon evolutionary preferences that may no longer reflect leadership ability (Klofstad & Anderson, 2018; Van Vugt et al., 2008b). In this case, it is possible that stereotypes about race with regards to dominance and aggressiveness affected the leadership ratings. However, our assumptions for this explanation, where Black men would be rated higher on perceived dominance and participants would prefer a Black man holding a leadership position (i.e., boss), did not hold. Overall, our exploratory analyses suggest that there is a possibility that the main effect of race upon leadership may be attributed to contrast effects or social demand effects, but we would need to conduct further research to determine the underlying mechanisms for these results.

In future research, we intend to address several limitations in our methodology that may have affected our results. Specifically, we only used White male voices, most of whom were graduate students at the University of Pennsylvania, which allowed us to have a relatively homogeneous sample of stimuli. Although race cannot be detected from the voice, SES and education levels are reflected by vocal characteristics, so it is entirely possible that participants were expecting the stereotypical Black names (which tend to be associated with low SES) to have African American vernacular (Labov, 2010). However, the voices we used as stimuli were from individuals that were attaining a much higher standard of education (PhD students) compared to the average individual, which may be obvious in their speech patterns. Additionally, the threat item was not situated in any context (i.e., participants were not provided any background as to why the voices should be

perceived as threatening), and it is possible that the ratings reflected different forms of perceived threat (i.e., physical threat, threat to resources, etc.). We observed the expected effect of voice pitch upon threat, which may reflect an innate understanding of how sounds can convey threat potential, which is observed even in infancy (e.g., larger objects produce a lower pitch) (Vestergaard et al., 2009). On the other hand, race is not an evolutionarily-relevant coalitional cue, but instead is constructed as a cue of coalitional alliances through ecological conditions (Kurzban & Leary, 2001), so it is unlikely that the manipulation of race elicited the perceptions of threat to the same degree as the voice manipulations. Since the study was conducted online and approximately half (49.5%) of the participants indicated that they listened to the recordings through speakers, it is also possible that there may have been differences in the listening environment that prevented participants from picking up on the differences in our voice pitch manipulations. Finally, we did not ask participants whether they thought the study was real, which may have provided more information about our observed results, since it is entirely possible that participants answered the suspicion check to align with the cover story because they thought it might have been an attention check.

It is imperative that future research within this domain attempts to address some of these limitations and determine whether the results are generalizable and replicable. Future studies should recruit a more diverse sample for vocal stimuli, including women and people from different racial groups and education levels. It would be useful to ask participants to guess which race and education level each voice represents to determine whether these characteristics will moderate the relationship between the independent variables and perceived leadership. Other studies (Hester & Gray, 2018) have included endorsement of stereotypes as a moderator for explaining higher ratings on stereotype-consistent items, which would be valuable in future extensions of this research. Finally, reversing the playback of the recordings may allow us to reduce the effects of speech content and vernacular upon participants' ratings.

It will also be important to disentangle the possible explanations for our unexpected effect of race upon perceived leadership. Specifically, researchers can overcome social demand effects by offering to pay participants to tell the truth. Future studies should use more objective measures of leadership, since contrast effects are more likely to appear when participants are rating stimuli on subjective measures (e.g., Likert-type items) because these ratings may vary across contexts, while objective measures are consistent, regardless of the target, the perceiver, or immediate environmental influences (Biernat, 2003). If future studies replicate the current study design but replace the leadership composite with objective measures and do not find a similar effect of race upon leadership, this will provide support for contrast effects upon our results.

Future extensions of this research will be fruitful in helping us fully understand the complex interplay of vocal characteristics and racial stereotypes in affecting person perception. As other research has demonstrated, Black men are more likely to be perceived as threatening (Trawalter, Todd, Baird, & Richeson, 2008; Wilson et al., 2017), which can be detrimental to their success in leadership positions when they are typically the minority in the corporate world. There is preliminary evidence in support of the concept that Black men that have disarming mechanisms may benefit from these personal characteristics in leadership positions (Hester & Gray, 2018; Livingston & Pearce, 2009; Wilson et al., 2017). Although the current study did not find the expected interaction effects of voice pitch and race upon perceived threat and leadership, there is still room for improvement in the methodological design and we encourage future researchers to extend this line of work to explore possible explanations further.

This line of work is important in helping us disentangle the personal characteristics that have a major effect upon person perception. Since we are incapable of reading others' minds to assess their intentions, we usually must make judgments of their character based upon their personal characteristics, even if these traits may not always be linked to their trustworthiness and/or threat potential. In this way, stereotypes about their group (based

upon their observable characteristics that cue group membership) and their personal characteristics that cue their ability to act upon any threatening intentions combine to predict trust towards that person.

The stereotype that Black people are a threat to physical safety and personal property has permeated largely because race and ecology are confounded in the United States, such that Black people are more likely to be impoverished, which is intricately linked with crime risk (Williams, Yu, Jackson, & Anderson, 1997; Williams, Sng, & Neuberg, 2016). Along these lines, Black people are overrepresented in certain contexts that link them with threat to physical safety (e.g., prisons) (Mauer & King, 2007; Roberts, 2004). Over time, Americans began to associate crime with any individuals categorized as Black (Quillian & Pager, 2001). These stereotypes are especially likely to be spread in the modern context, since there are a multitude of technologies available for communicating with numerous individuals regardless of interpersonal distance, which encourages the uniform and rapid spread of information across a culture. Also, popular press facilitates this stereotyping by reporting racialized crime stories that strengthen the association between race and physical threat (Dixon & Azocar, 2007; Gilliam, Iyengar, Simon, & Wright, 1996). Through these mechanisms, stereotypes have become ingrained in the public conscious in America today, and continue to affect how Black people are treated daily, even after explicit prejudice has become less socially acceptable (Murphy, Richeson, Shelton, Rheinschmidt, & Bergsieker, 2013). Since many of the stereotypes in America are rooted in protracted racial tensions throughout history, any intervention to reduce these stereotypes must be comprehensive, targeting the many factors that contribute to stereotypes about Black people. Specifically, it is possible that the voice may serve as a potent disarming mechanism that can reduce perceived threat in Black men, which we explored through the current study. This research provides an initial glimpse into how certain vocal characteristics can be affected by racial stereotypes, but future research needs to continue this line of work to enlighten us about the importance of nonverbal behavior in

838 influencing perceptions of, and in turn, behavior towards minority individuals.

References

Apicella, C. L., & Feinberg, D. R. (2009). Voice pitch alters mate-choice-relevant perception in hunter-gatherers. *Proceedings of the Royal Society B: Biological Sciences*, *276*, 1077–1082. https://doi.org/10.1098/rspb.2008.1542

Ayres, I., Banaji, M., & Jolls, C. (2015). Race effects on eBay. *RAND Journal of Economics*, *46*(4), 891–917. https://doi.org/10.1111/1756-2171.12115

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Barr, D., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 1–43. https://doi.org/10.1016/j.jml.2012.11.001.Random

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. Retrieved from http://arxiv.org/abs/1506.04967

Biernat, M. (2003). Toward a broader view of social stereotyping. *American Psychologist*, *58*(12), 1019–1027. https://doi.org/10.1037/0003-066X.58.12.1019

Biernat, M., Manis, M., & Nelson, T. E. (1991). Stereotypes and standards of judgment. *Journal of Personality and Social Psychology*, *60*(4), 485–499. https://doi.org/10.1037/0022-3514.60.4.485

Boersma, P., & Heuven, V. van. (2001). Speak and unSpeak with Praat. *Glot International*, *5*(9-10), 341–347.

Bowles, S. (2009). Did warfare among ancestral hunter-gatherer affect the evolution of human social behaviors? *Science*, *324*, 1293–1298. Retrieved from http://science.sciencemag.org/content/324/5932/1293.short

Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods*, *23*(3), 389–411.

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, *1*(3), 473–514. https://doi.org/10.1214/06-BA117

Clarke, P., & Wheaton, B. (2007). Addressing data sparseness in contextual population research using cluster analysis to create synthetic neighborhoods. *Sociological Methods & Research*, *35*(3), 311–351.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2007). The influence of stereotypes on decisions to shoot. *European Journal of Social Psychology*, *37*, 1102–1117. https://doi.org/10.1002/ejsp

Cottrell, C. A., & Neuberg, S. L. (2005). Different emotional reactions to different groups: A sociofunctional threat-based approach to "prejudice". *Journal of Personality and Social Psychology*, *88*(5), 770–789. https://doi.org/10.1037/0022-3514.88.5.770

Devine, P. G., & Elliot, A. J. (1995). Are Racial Stereotypes Really Fading? The Princeton Trilogy Revisited. *Personality and Social Psychology Bulletin*, *21*(11), 1139–1150.

Dixon, T. L., & Azocar, C. L. (2007). Priming crime and activating blackness: Understanding the psychological impact of the overrepresentation of blacks as lawbreakers on television news. *Journal of Communication*, *57*(2), 229–253. https://doi.org/10.1111/j.1460-2466.2007.00341.x

Doleac, J. L., & Stein, L. C. D. (2013). The visible hand: Race and online market outcomes. *The Economic Journal*, *123*(572), 1–18. https://doi.org/10.1111/ecoj.12082

888 Eberhardt, J. L., Davies, P. G., Purdie-Vaughns, V. J., & Johnson, S. L. (2006). Looking
889 deathworthy perceived stereotypicality of black defendants predicts
890 capital-sentencing outcomes. *Psychological Science*, *17*(5), 383–386.
891 https://doi.org/10.1111/j.1467-9280.2006.01716.x

892 Elff, M., Heisig, J. P., Schaeffer, M., & Shikano, S. (2020). Multilevel analysis with few
893 clusters: Improving likelihood-based methods to provide unbiased estimates and
894 accurate inference. *British Journal of Political Science*, 1–15.
895 https://doi.org/10.1017/S0007123419000097

896 Fairbanks, G. (1960). *Voice and articulation drillbook*. New York, NY: Harper.

897 Feinberg, D. R., Jones, B. C., Little, A. C., Burt, D. M., & Perrett, D. I. (2005).
898 Manipulations of fundamental and formant frequencies influence the attractiveness
899 of human male voices. *Animal Behaviour*, *69*(3), 561–568.
900 https://doi.org/10.1016/j.anbehav.2004.06.012

901 Finch, W. H., Bolin, J. E., & Kelley, K. (2014). *Multilevel modeling using R*.

902 Fraccaro, P. J., Jones, B. C., Vukovic, J., Smith, F. G., Watkins, C. D., Feinberg, D. R., ...
903 Debruine, L. M. (2011). Experimental evidence that women speak in a higher voice
904 pitch to men they find attractive. *Journal of Evolutionary Psychology*, *9*(1), 57–67.
905 https://doi.org/10.1556/JEP.9.2011.33.1

906 Fraccaro, P. J., O'Connor, J. J. M., Re, D. E., Jones, B. C., DeBruine, L. M., & Feinberg,
907 D. R. (2013). Faking it: Deliberately altered voice pitch and vocal attractiveness.
908 *Animal Behaviour*, *85*(1), 127–136. https://doi.org/10.1016/j.anbehav.2012.10.016

909 Gaddis, S. (2017). How Black are Lakisha and Jamal? Racial perceptions from names used
910 in correspondence audit studies. *Sociological Science*, *4*, 469–489.
911 https://doi.org/10.15195/v4.a19

912 Garson, G. D. (2013). *Hierarchical linear modeling: Guide and applications*.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*.

Gilliam, F. D., Iyengar, S., Simon, A., & Wright, O. (1996). Crime in black and white: The violent, scary world of local news. *Harvard International Journal of Press/Politics*, *1*(3), 6–23. https://doi.org/10.1177/1081180X96001003003

Harries, M., Hawkins, S., Hacking, J., & Hughes, I. (1998). Changes in the male voice at puberty: vocal fold length and its relationship to the fundamental frequency of the voice. *The Journal of Laryngology and Otology*, *112*, 451–454.

Hayward, M. D., Miles, T. P., Crimmins, E. M., & Yang, Y. (2000). The significance of socioeconomic status in explaining the racial gap in chronic health conditions. *American Sociological Review*, *65*(6), 910–930.

Hester, N., & Gray, K. (2018). For Black men, being tall increases threat stereotyping and police stops. *Proceedings of the National Academy of Sciences*, *115*(11), 2711–2715. https://doi.org/10.1073/pnas.1714454115

Hodges-Simeon, C. R., Gurven, M., & Gaulin, S. J. C. (2015). The low male voice is a costly signal of phenotypic quality among Bolivian adolescents. *Evolution and Human Behavior*, *36*(4), 1–9. https://doi.org/10.1016/j.evolhumbehav.2015.01.002

Hodges-Simeon, C. R., Gurven, M., Puts, D., & Gaulin, S. (2014). Vocal fundamental and formant frequencies are honest signals of threat potential in peripubertal males. *Behavioral Ecology*, *25*(4), 984–988. https://doi.org/10.1093/beheco/aru081

Hughes, S. M., Mogilski, J. K., & Harrison, M. A. (2014). The perception and parameters of intentional voice manipulation. *Journal of Nonverbal Behavior*, *38*(1), 107–127. https://doi.org/10.1007/s10919-013-0163-z

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of*

Psychology, *68*, 17.1–17.25. https://doi.org/10.1146/annurev-psych-122414-033702

Kirkpatrick, S. A., & Locke, E. A. (1991). Leadership: Do traits matter? *The Executive*, *5*(2), 48–60. https://doi.org/10.5465/AME.1991.4274679

Klofstad, C. A., & Anderson, R. C. (2018). Voice pitch predicts electability , but does not signal leadership ability. *Evolution and Human Behavior*. https://doi.org/10.1016/j.evolhumbehav.2018.02.007

Klofstad, C. A., Anderson, R. C., & Peters, S. (2012). Sounds like a winner: Voice pitch influences perception of leadership capacity in both men and women. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1738), 2698–2704. https://doi.org/10.1098/rspb.2012.0311

Knuycky, L. R., Kleider, H. M., & Cavrak, S. E. (2014). Line-up misidentifications: When being 'prototypically black' is perceived as criminal. *Applied Cognitive Psychology*, *28*(1), 39–46. https://doi.org/10.1002/acp.2954

Kreiman, J., & Sidtis, D. (2011). *Foundations of voice studies: An interdisciplinary approach to voice production and perception*.

Kurzban, R., & Leary, M. R. (2001). Evolutionary origins of stigmatization: The functions of social exclusion. *Psychological Bulletin*, *127*(2), 187–208. https://doi.org/10.1037/0033-2909.127.2.187

Labov, W. (2010). Unendangered dialect, endangered people: The case of African American vernacular English. *Transforming Anthropology*, *18*(1), 15–28. https://doi.org/10.1111/j.1548-7466.2010.01066.x.15

Li, N. P., Vugt, M. van, & Colarelli, S. M. (2017). The Evolutionary Mismatch Hypothesis: Implications for Psychological Science. *Current Directions in Psychological Science*, 096372141773137. https://doi.org/10.1177/0963721417731378

Livingston, R. W., & Pearce, N. A. (2009). The teddy-bear effect: Does having a baby face

963 benefit black chief executive officers? *Psychological Science*, *20*(10), 1229–1236.
964 https://doi.org/10.1111/j.1467-9280.2009.02431.x

965 Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior*
966 *Research Methods*, *49*, 1494–1502. https://doi.org/10.3758/s13428-016-0809-y

967 Maas, C. J. M., & Hox, J. J. (2004). The influence of violations of assumptions on
968 multilevel parameter estimates and their standard errors. *Computational Statistics*
969 *and Data Analysis*, *46*(3), 427–440. https://doi.org/10.1016/j.csda.2003.08.006

970 Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling.
971 *Methodology*, *1*(3), 86–92. https://doi.org/10.1027/1614-1881.1.3.86

972 Maner, J. K., Kenrick, D. T., Backer, D. V., Robertson, T. E., Hofer, B., Neuberg, S. L.,
973 ... Schaller, M. (2005). Functional projection: How fundamental social motives can
974 bias interpersonal perception. *Journal of Personality and Social Psychology*, *88*(1),
975 63–78. https://doi.org/10.1037/0022-3514.88.1.63

976 Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I
977 error and power in linear mixed models. *Journal of Memory and Language*, *94*,
978 305–315. https://doi.org/10.1016/j.jml.2017.01.001

979 Mauer, M., & King, R. S. (2007). Uneven justice: State rates of incarceration by race and
980 ethnicity. *The Sentencing Project*, 1–23. Retrieved from
981 http://sites.google.com/site/lkeber/SentencingProjRatesofIncarcerationby.pdf

982 Mayew, W. J., Parsons, C. A., & Venkatachalam, M. (2013). Voice pitch and the labor
983 market success of male chief executive officers. *Evolution and Human Behavior*,
984 *34*(4), 243–248. https://doi.org/10.1016/j.evolhumbehav.2013.03.001

985 McNeish, D. (2017). Small sample methods for multilevel modeling: A colloquial
986 elucidation of REML and the Kenward-Roger correction. *Multivariate Behavioral*
987 *Research*, *52*(5), 661–670. https://doi.org/10.1080/00273171.2017.1344538

McNeish, D., & Stapleton, L. M. (2016). Modeling clustered data with very few clusters. *Multivariate Behavioral Research*, *51*(4), 495–518. https://doi.org/10.1080/00273171.2016.1167008

Meteyard, L., & Davies, R. A. (2020). Best practice guidance for LMMs. *Journal of Memory and Language*, *112*, 1–62. https://doi.org/DOI%2010.17605/OSF.IO/BFQ39

Misangyi, V. F., LePine, J. A., Algina, J., & Geoddeke Jr., F. (2006). The adequacy of repeated-measures regression for multilevel research: Comparisons with repeated-measures ANOVA, multivariate repeated-measures ANOVA, and multilevel modeling across various multilevel research designs. *Organizational Research Methods*, *9*(1), 5–28.

Murphy, M. C., Richeson, J. A., Shelton, J. N., Rheinschmidt, M. L., & Bergsieker, H. B. (2013). Cognitive costs of contemporary prejudice. *Group Processes and Intergroup Relations*, *16*(5), 560–571. https://doi.org/10.1177/1368430212468170

Navarrete, C. D., McDonald, M. M., Molina, L. E., & Sidanius, J. (2010). Prejudice at the nexus of race and gender: An outgroup male target hypothesis. *Journal of Personality and Social Psychology*, *98*(6), 933–945. https://doi.org/10.1037/a0017931

Neuberg, S. L., & Cottrell, C. A. (2008). Managing the threats and opportunities afforded by human sociality. *Group Dynamics*, *12*(1), 63–72. https://doi.org/10.1037/1089-2699.12.1.63

Neuberg, S. L., & Schaller, M. (2016). An evolutionary threat-management approach to prejudices. *Current Opinion in Psychology*, *7*, 1–5. https://doi.org/10.1016/j.copsyc.2015.06.004

O'Connor, J. J. M., & Barclay, P. (2017). The influence of voice pitch on perceptions of trustworthiness across social contexts. *Evolution and Human Behavior*, *38*(4),

506–512. https://doi.org/10.1016/j.evolhumbehav.2017.03.001

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087–11092. https://doi.org/10.1073/pnas.0805664105

Pisanski, K., Mora, E. C., Pisanski, A., Reby, D., Sorokowski, P., Frackowiak, T., & Feinberg, D. R. (2016). Volitional exaggeration of body size through fundamental and formant frequency modulation in humans. *Scientific Reports*, *6*, 1–8. https://doi.org/10.1038/srep34389

Puts, D. A. (2010). Beauty and the beast: Mechanisms of sexual selection in humans. *Evolution and Human Behavior*, *31*(3), 157–175. https://doi.org/10.1016/j.evolhumbehav.2010.02.005

Puts, D. A., Apicella, C. L., & Cardenas, R. A. (2012). Masculine voices signal men's threat potential in forager and industrial societies. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1728), 601–609. https://doi.org/10.1098/rspb.2011.0829

Puts, D. A., Gaulin, S. J. C., & Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism in human voice pitch. *Evolution and Human Behavior*, *27*(4), 283–296. https://doi.org/10.1016/j.evolhumbehav.2005.11.003

Quillian, L., & Pager, D. (2001). Black neighbors, higher crime? The role of racial stereotypes in evaluations of neighborhood crime. *American Journal of Sociology*, *107*(3), 717–767. https://doi.org/10.1086/338938

Raudenbush, S., & Byrk, A. (2002). *Hierarchical linear models: Applications and data analysis methods*.

Riach, P., & Rich, J. (2002). Field experiments of discrimination in the market place. *The Economic Journal*, *112*(483), F480–F518.

Roberts, D. E. (2004). The social and moral cost of mass incarceration in African

American communities. *Stanford Law Review*, *56*(5), 1271–1305. https://doi.org/10.2307/40040178

Rosette, A. S., Leonardelli, G. J., & Phillips, K. W. (2008). The White standard: Racial bias in leader categorization. *Journal of Applied Psychology*, *93*(4), 758–777. https://doi.org/10.1037/0021-9010.93.4.758

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika*, *6*(5), 309–316.

Scherbaum, C. A., & Ferreter, J. M. (2009). Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organizational Research Methods*, *12*(2), 347–367.

Schielzeth, H., & Forstmeier, W. (2009). Conclusions beyond support: Overconfident estimates in mixed models. *Behavioral Ecology*, *20*, 416–420. https://doi.org/10.1093/beheco/arn145

Sidanius, J., & Veniegas, R. C. (2000). Gender and race discrimination: The interactive nature of disadvantage. *Reducing Prejudice and Discrimination*, 47–69.

Smid, S. C., Mcneish, D., Miočević, M., & Schoot, R. V. D. (2020). Bayesian versus frequentist estimation for structural equation models in small sample contexts: A systematic review. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*, 131–161. https://doi.org/10.1080/10705511.2019.1577140

Stanley, D. A., Sokol-Hessner, P., Banaji, M. R., & Phelps, E. A. (2011). Implicit race attitudes predict trustworthiness judgments and economic trust decisions. *Proceedings of the National Academy of Sciences*, *108*(19), 7710–7715. https://doi.org/10.1073/pnas.1014345108

Stanley, D. A., Sokol-Hessner, P., Fareri, D. S., Perino, M. T., Delgado, M. R., Banaji, M. R., & Phelps, E. A. (2012). Race and reputation: Perceived racial group trustworthiness influences the neural correlates of trust decisions. *Philosophical*

Transactions of the Royal Society B: Biological Sciences, *367*(1589), 744–753. https://doi.org/10.1098/rstb.2011.0300

Stegmueller, D. (2013). How many countries for multilevel modeling? A comparison of frequentist and bayesian approaches. *American Journal of Political Science*, *57*(3), 748–761. https://doi.org/10.1111/ajps.12001

Tigue, C. C., Borak, D. J., O'Connor, J. J. M., Schandl, C., & Feinberg, D. R. (2012). Voice pitch influences voting behavior. *Evolution and Human Behavior*, *33*(3), 210–216. https://doi.org/10.1016/j.evolhumbehav.2011.09.004

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Unconscious evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*(12), 455–460. https://doi.org/10.1037/a0027950

Trawalter, S., Todd, A. R., Baird, A. A., & Richeson, J. A. (2008). Attending to threat: Race-based patterns of selective attention. *Journal of Experimental Social Psychology*, *44*(5), 1322–1327. https://doi.org/10.1016/j.jesp.2008.03.006

Van Vugt, M., Hogan, R., & Kaiser, R. B. (2008a). Leadership, followership, and evolution. *American Psychologist*, *63*(3), 182–196. https://doi.org/10.1037/0003-066X.63.3.182

Van Vugt, M., Johnson, D. D. P., Kaiser, R. B., & O'Gorman, R. (2008b). Evolution and the social psychology of leadership: The mismatch hypothesis. In *Social psychology and leadership* (pp. 1–26). https://doi.org/10.1336/027599760x

Vestergaard, M. D., Háden, G. P., Shtyrov, Y., Patterson, R. D., Pulvermüller, F., Denham, S. L., … Winkler, I. (2009). Auditory size-deviant detection in adults and newborn infants. *Biological Psychology*, *82*(2), 169–175. https://doi.org/10.1016/j.biopsycho.2009.07.004.Auditory

Voelter, C., Kleinsasser, N., Joa, P., Nowack, I., Martínez, R., Hagen, R., & Voelker, H. (2008). Detection of hormone receptors in the human vocal fold. *European Archives*

of Otorhinolaryngol, *265*, 1239–1244. https://doi.org/10.1007/s00405-008-0632-x

Vogel, A. P., Maruff, P., Snyder, P. J., & Mundt, J. C. (2009). Standardization of pitch range settings in voice acoustic analysis. *Behavior Research Methods*, *41*(2), 318–324. https://doi.org/10.3758/BRM.41.2.318.Standardization

Vukovic, J., Jones, B. C., Debruine, L., Feinberg, D. R., Smith, F. G., Little, A. C., ... Main, J. (2010). Women's own voice pitch predicts their preferences for masculinity in men's voices. *Behavioral Ecology*, *21*(4), 767–772. https://doi.org/10.1093/beheco/arq051

Welch, K. (2007). Black criminal stereotypes and racial profiling. *Journal of Contemporary Criminal Justice*, *23*(3), 276–288.

Westfall, J., Judd, C. M., & Kenny, D. A. (2015). Replicating studies in which samples of participants respond to samples of stimuli. *Perspectives on Psychological Science*, *10*(3), 390–399. https://doi.org/10.1177/1745691614564879

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 1–26. https://doi.org/10.1037/xge0000014

Williams, D. R., Yu, Y., Jackson, J. S., & Anderson, N. B. (1997). Racial differences in physical and mental health: Socio-economic status, stress and discrimination. *Journal of Health Psychology*, *2*(3), 335–351. https://doi.org/10.1177/135910539700200305

Williams, K. E. G., Sng, O., & Neuberg, S. L. (2016). Ecology-driven stereotypes override race stereotypes. *Proceedings of the National Academy of Sciences*, *113*(2), 310–315. https://doi.org/10.1073/pnas.1519401113

Wilson, J. P., Hugenberg, K., & Rule, N. O. (2017). Racial bias in judgments of physical

size and formidability: From size to threat. *Journal of Personality and Social Psychology*, *113*(1), 59–80. https://doi.org/10.1037/pspi0000092.supp